Position: Carbon Footprint Reporting Should Be Routine in Machine Learning Research

Anonymous Authors[1]

Abstract

In this position paper, we argue that the machine learning community should adopt standardized carbon footprint reporting as part of routine scientific practice. Training large models can emit hundreds of tons of CO_2, yet environmental costs remain largely invisible in publications. We contend that without energy and emissions metrics, claims of model efficiency are incomplete: a method cannot be deemed "efficient" without specifying efficient at what. This gap undermines scientific rigor and reproducibility, as identical experiments in different locations yield vastly different carbon footprints. We put forth reporting guidelines comprising five standardized metrics, practical measurement tools, and integration with community benchmarks, with a phased three-stage adoption process. We address alternative views, including concerns about measurement complexity and potential barriers for resource-limited researchers. To promote equity, we advocate for dual reporting of energy and carbon, reference-grid normalization, and acceptance of approximate estimates. This paper calls on venues, reviewers, authors, and institutions to establish carbon awareness as a foundational element of responsible ML research.

1. Introduction

Machine learning progress has a cost our community rarely quantifies: the energy consumed and carbon emitted during training and deployment. A typical ML paper specifies GPU types, training duration, and hyperparameters in detail, yet omits energy consumption or carbon emissions entirely. This asymmetry reflects a blind spot in our scientific culture.

The environmental impact is substantial. Training a Transformer model with neural architecture search can emit as much carbon as five cars over their lifetimes (Strubell et al., 2019). Computational requirements have increased roughly 10× every two years (Sevilla et al., 2022), and data centers now consume 1–2% of global electricity (Masanet et al., 2020; Xiao et al., 2025).

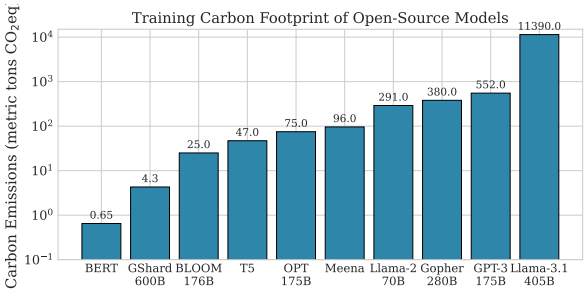

Figure 1. Officially published carbon emissions for training major language models (training-phase only).[1] Variation reflects model scale and infrastructure: BLOOM (25 tons) used low-carbon French grid; Llama-3.1 405B is the largest disclosed at 11,390 tons. Methodologies vary; direct comparisons require caution.

As shown in Figure 1, emissions vary dramatically: GPT-3 emitted 552 tons (Patterson et al., 2021), while BLOOM achieved comparable scale at just 25 tons using France's nuclear-powered grid (Luccioni et al., 2023). This variation demonstrates how much infrastructure choices matter, yet most model announcements omit environmental costs entirely.

We argue that **standardized carbon footprint reporting should become a routine component of ML research publications**. Without energy and emissions metrics, efficiency claims lack scientific rigor.

This is not environmental activism but scientific completeness. Physicists report energy in collision experiments; chemists report temperature in synthesis protocols. ML researchers claiming efficient methods should report energy consumption. Our position rests on four observations: (1) ef-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]Data sources: BERT (Strubell et al., 2019); T5, Meena, GShard-600B, GPT-3 (Patterson et al., 2021); OPT-175B (Zhang et al., 2022); BLOOM, Gopher (Luccioni et al., 2023); Llama-2 (Touvron et al., 2023b); Llama-3.1 (Meta AI, 2024).

ficiency claims require specifying efficient *at what* (FLOPs, energy, carbon, or dollars); (2) reproducibility demands environmental context, as identical experiments in different locations yield vastly different carbon footprints; (3) the field lacks baselines to track sustainability over time; and (4) awareness plausibly drives change—while direct empirical evidence in ML is limited, analogous norms in other fields (code release, reproducibility checklists) suggest that what gets measured gets improved.

Now is the time to act. Measurement tools exist, researcher awareness is growing, and prominent releases including BLOOM, OPT, and Llama have already included carbon data, proving such reporting is feasible.

2. Related Work

Green computing research has studied data center efficiency for decades (Masanet et al., 2020). Within ML, Schwartz et al. introduced "Green AI" (Schwartz et al., 2020), and Strubell et al. brought attention to NLP's carbon costs (Strubell et al., 2019). Patterson et al. provided refined estimates for large-scale training (Patterson et al., 2021), while Luccioni et al. documented BLOOM's carbon footprint (Luccioni et al., 2023). Lottick et al. framed energy reporting as algorithmic accountability (Lottick et al., 2019), and recent work extends carbon analysis to adversarial ML and security domains (Hasan et al., 2024). K.C. et al. demonstrate that per-experiment carbon tracking is feasible beyond NLP and vision, integrating real-time CodeCarbon monitoring into cybersecurity anomaly detection workflows (K.C. et al., 2025). Henderson et al. found that the vast majority of papers at major venues omit energy data entirely (Henderson et al., 2020).

Our proposal draws on precedents from other fields: climate science frameworks for reporting uncertainty, medical research's CONSORT standards,[2] software engineering norms for code sharing, and prior work on structured reporting in ML (Dodge et al., 2019). Other scientific disciplines have begun addressing research carbon footprints, including astronomy (Knödlseder et al., 2022).

The scaling laws literature provides context for carbon costs. Kaplan et al. established performance power laws with respect to compute (Kaplan et al., 2020), and Hoffmann et al. showed that training smaller models on more data can achieve equivalent performance at lower cost (Hoffmann et al., 2022). Recent work has expanded understanding of ML's footprint: MLPerf Power[3] established standardized efficiency measurement methodology (Tschand et al., 2025); studies show inference energy can exceed training costs for deployed models (Jegham et al., 2025; Fernandez et al., 2025); analyses project substantial increases in AI electricity demand (Xiao et al., 2025); and green AI techniques demonstrate 40–60% energy reductions without performance degradation (Verdecchia et al., 2025).

Several recent efforts have developed carbon-aware ML systems: OpenCarbonEval provides unified emission estimation (Yu et al., 2024); Clover enables carbon-aware inference routing (Li et al., 2023); and CAFE addresses carbon-aware federated learning across distributed data centers (Bian et al., 2023). We argue these technical advances should be complemented by venue-level reporting standards, and offer concrete metrics, templates, and a phased adoption timeline toward this goal.

3. The Current State of Carbon Reporting in ML

Despite growing awareness of computational sustainability, carbon footprint reporting remains rare in ML publications. Henderson et al. conducted a systematic survey of 100 randomly sampled NeurIPS 2019 papers and found striking results: zero papers reported carbon impacts, only 1% reported any energy metrics, and just 17% reported compute-related metrics such as GPU-hours (Henderson et al., 2020). While awareness has grown since 2019, reporting remains the exception rather than the rule.

3.1. What Gets Reported

The standard computational details section of an ML paper typically includes hardware specifications such as GPU model, memory, and number of devices; training duration in wall-clock time or epochs; hyperparameters and optimization details; and dataset sizes with preprocessing steps. These details serve reproducibility but say nothing about environmental impact. A researcher reading that a model trained for 72 hours on 8 A100 GPUs cannot determine whether the experiment emitted 50 kg or 500 kg of CO_2 without additional information about data center location and energy sources.

3.2. Why Reporting Matters for Science

The absence of energy metrics creates several problems for scientific practice:

Incomplete efficiency claims. When a paper claims a model is "more efficient," this typically means fewer FLOPs or faster inference. But FLOPs do not map linearly to energy consumption. Memory access patterns, hardware utilization, and batch sizes all affect energy use independently of operation counts. A model with 20% fewer FLOPs might

[2] Consolidated Standards of Reporting Trials, a guideline for reporting randomized controlled trials in medicine.

[3] An industry benchmark suite for measuring ML system performance and energy efficiency.

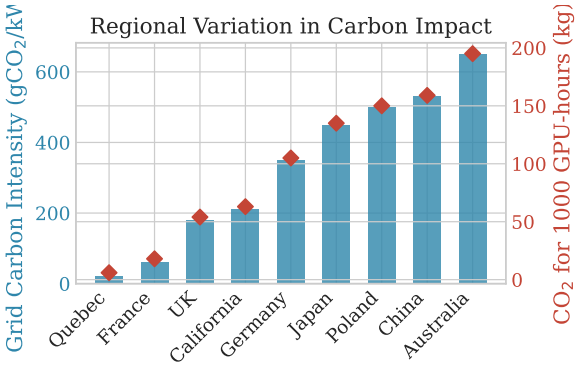

Figure 2. Regional variation in grid carbon intensity and resulting emissions for 1000 GPU-hours, ranging from 6 kg in Quebec to 195 kg in Australia.

consume the same energy if it has worse memory locality.

Non-reproducible comparisons. Two research groups comparing methods on identical hardware may get different energy results depending on their locations. A GPU cluster in Quebec (hydroelectric grid, approximately 20 gCO_2/kWh) produces roughly 25× less carbon than one in Poland (coal-heavy grid, approximately 500 gCO_2/kWh) for the same computation (Lacoste et al., 2019). Without location data, carbon comparisons across papers become uninterpretable. Energy comparisons (kWh) and algorithmic efficiency metrics (FLOPs per token) remain valid without location data, but carbon footprint claims require this context.

Hidden costs of progress. When we celebrate a new state-of-the-art result, we rarely ask what it cost to achieve. The computational experiments behind a single paper can range from tens of GPU-hours to millions. Without systematic reporting, we cannot assess whether marginal accuracy gains justify orders-of-magnitude increases in compute.

Lack of optimization incentives. When energy costs are invisible, researchers have no incentive to optimize for them. A training run that uses 2× more energy than necessary due to suboptimal batch sizes or learning rate schedules will produce the same paper as an efficient run. If energy were reported, reviewers and readers could recognize and appreciate efficient implementations, creating positive incentives for carbon-conscious research practices.

3.3. A Worked Example

Consider two research groups training identical models with 4 V100 GPUs for 24 hours. Group A operates in Norway (29 gCO_2/kWh, PUE 1.1); Group B in Australia (650 gCO_2/kWh, PUE 1.4). Both report identical GPU-hours, yet emissions differ by 25×: approximately 1 vs 25 $kgCO_2$eq. Without location data, these experiments appear equivalent. Figure 2 visualizes this variation.

3.4. Existing Tools and Their Limitations

Several tools exist for measuring ML carbon footprints: CodeCarbon (CodeCarbon Contributors, 2020), ML CO2 Impact[4] (Lacoste et al., 2019), Carbontracker (Anthony et al., 2020), and experiment-impact-tracker (Henderson et al., 2020). These tools can estimate energy consumption and convert it to carbon emissions using regional grid intensity data.

However, tool availability has not led to widespread adoption, primarily because venues do not encourage or reward carbon reporting.

3.5. The Gap Between Awareness and Action

The ML community is aware of environmental concerns: workshops on climate change and AI regularly draw submissions, and papers on efficient methods often mention environmental benefits. Yet this awareness rarely translates into consistent reporting.

Several factors explain this gap. First, there is no template. Authors who want to report carbon footprint must decide what to measure, how to measure it, and how to present results. Without guidance, many default to reporting nothing. Second, there is no enforcement. Reviewers do not expect carbon data, so its absence is not penalized. Third, there is uncertainty about accuracy. Researchers worry that imprecise estimates will be criticized, so they prefer to omit data rather than report uncertain figures.

This pattern is not unique to environmental reporting. The field has faced similar challenges with statistical practices (e.g., reporting confidence intervals), code release (now expected), and dataset documentation. In each case, progress required venues to establish clear expectations and provide templates that make compliance straightforward.

4. Proposed Reporting Guidelines

The gap between awareness and action described above will not close on its own. Without concrete standards, authors face uncertainty about what to report and how, leading most to report nothing. Standardization reduces this friction by providing clear templates and shared expectations.

We propose standardized reporting guidelines for carbon footprint in ML research. The goal is straightforward: every paper should report the energy consumed and carbon emitted by its experiments, using consistent metrics that enable comparison across studies. Our guidelines prioritize

[4]ML CO2 Impact calculator: https://mlco2.github.io/impact/

practicality over perfection—estimates with acknowledged uncertainty are preferable to no data at all.

4.1. Standardized Metrics for Reporting

The following metrics align with established carbon accounting standards: the Greenhouse Gas Protocol (World Resources Institute and World Business Council for Sustainable Development, 2004) for organizational carbon accounting[5] and the Software Carbon Intensity (SCI) specification (Green Software Foundation, 2022) for software applications.

Building on these frameworks, papers should report five key metrics where applicable:

1. **Total energy consumption (kWh):** Direct measurement via hardware power meters provides the most accurate values. When unavailable, software-based tools such as CodeCarbon, Carbontracker, or nvidia-smi can provide reasonable estimates, though researchers should acknowledge measurement uncertainty (Jay et al., 2023). For cloud workloads, providers increasingly offer energy consumption data through APIs and dashboards.

2. **Carbon emissions ($kgCO_2eq$):** Calculated by multiplying energy consumption by grid carbon intensity, following the methodology outlined in the GHG Protocol. The unit $kgCO_2eq$ (kilograms of carbon dioxide equivalent) accounts for all greenhouse gases converted to their CO_2 warming potential.

3. **Grid carbon intensity (gCO_2/kWh):** This value varies dramatically by region, from approximately 20 gCO_2/kWh in hydroelectric-dominated Quebec to over 700 gCO_2/kWh in coal-heavy grids. Researchers should cite their data source, such as regional grid operators, Electricity Maps[6] (Electricity Maps, 2023), or annual averages from governmental statistics. Time-of-use variations can also be significant.

4. **Power Usage Effectiveness (PUE):** This ratio captures data center overhead including cooling, lighting, and power distribution losses. A PUE of 1.0 represents perfect efficiency; typical values range from 1.1 for hyperscale cloud data centers to 1.5 or higher for older academic facilities. When unknown, a conservative estimate of 1.2–1.4 should be stated explicitly.

5. **Compute region:** The geographic location (country, state, or cloud availability zone) enables readers to verify or recalculate carbon estimates and facilitates meta-analyses across studies.

Following the GHG Protocol's emissions categorization, ML training typically falls under Scope 2 (purchased electricity), while Scope 3 includes embodied carbon from hardware manufacturing and supply chain emissions (Gupta et al., 2022). For comprehensive lifecycle assessment, researchers conducting large-scale training should consider reporting embodied carbon estimates, though we recognize this requires additional data that may not always be available (Ligozat et al., 2022).

For cloud experiments, Google Cloud, Microsoft Azure, and AWS all offer carbon footprint dashboards that researchers can cite directly.

4.2. Measurement Tools and Carbon-Aware Benchmarks

Adoption of carbon reporting depends heavily on reducing friction. Several open-source tools already exist for this purpose, including CodeCarbon (CodeCarbon Contributors, 2020), Carbontracker (Anthony et al., 2020), and experiment-impact-tracker (Henderson et al., 2020). Software-based estimates generally track hardware measurements within 20–30% for GPU-intensive workloads, which is sufficient for the transparency goals of carbon reporting. Even with this uncertainty, cross-paper comparisons remain meaningful: a 10× difference between methods is clearly distinguishable from measurement noise, and trend analysis across hundreds of papers will average out individual errors.

Ideally, measurement should require minimal code changes. Decorators or context managers make tracking nearly invisible:

```
@track_emissions(project="training")
def train():
    model.fit(X, y, epochs=100)
```

Output formats should be standardized across tools, and HPC clusters should enable job-level energy accounting through schedulers like SLURM.[7]

Beyond paper reporting, community benchmarks should incorporate carbon metrics. MLPerf Power has established methodology for standardized efficiency benchmarking (Tschand et al., 2025). Leaderboards could display carbon cost per accuracy point, enabling identification of Pareto-optimal models. Green AI techniques can reduce energy consumption substantially (Verdecchia et al., 2025), but these gains remain invisible without carbon-aware evaluation.

[5] ISO 14064-1:2018 provides complementary international standards for emissions verification.

[6] https://electricitymaps.com

[7] Simple Linux Utility for Resource Management, a widely used workload manager for high-performance computing clusters.

4.3. Choice of Reporting Primitives

Different reporting units serve different purposes, and the community must choose which primitives to standardize. We consider several candidates and their tradeoffs:

Energy-based primitives. Metrics like kWh (total energy) or Wh/1000 tokens (energy intensity) measure physical resource consumption independent of location. These are reproducible and enable fair comparison of algorithmic efficiency across institutions. However, they do not capture environmental impact, which depends on carbon intensity.

Carbon-based primitives. Metrics like $kgCO_2eq$ (total emissions) or gCO_2/1000 tokens directly measure environmental impact. These are the ultimate quantity of interest for sustainability but depend on location and infrastructure, making cross-study comparison difficult without normalization.

Compute-normalized metrics. Metrics like kWh/billion parameters or gCO_2/accuracy-point enable comparison across models of different scales. These are useful for identifying efficient architectures but require careful definition of the denominator.

System vs. paper-level reporting. Benchmarks like MLPerf Power (Tschand et al., 2025) measure system efficiency under controlled conditions (fixed hardware, standardized workloads, comparable environments). Paper-level reporting captures the full cost of research including failed experiments and hyperparameter search. Both are valuable: system benchmarks enable hardware comparison while paper-level reporting provides scientific transparency.

We recommend a layered approach: papers should report *both* energy (kWh) and carbon ($kgCO_2eq$), along with the inputs needed to verify the calculation (grid intensity, PUE, region). This enables readers to compare energy efficiency directly while understanding the carbon implications. For inference, standardized benchmarks (e.g., Wh per 1000 tokens at batch size 32) complement paper-specific totals.

4.4. Reporting Training and Inference Separately

Training and inference have fundamentally different carbon characteristics and must be reported separately (Luccioni et al., 2024; Samsi et al., 2023; Wu et al., 2022; Jegham et al., 2025).

Training carbon is a one-time, upfront investment. It scales with model size following established scaling laws: larger models require more compute, and compute requirements grow polynomially with parameter count (Kaplan et al., 2020; Hoffmann et al., 2022). The Chinchilla scaling laws suggest that compute-optimal training involves more tokens rather than simply larger models (Hoffmann et al., 2022). Training can be scheduled strategically to exploit periods of low grid carbon intensity or high renewable availability (Dodge et al., 2022), and recent work on multi-day carbon intensity forecasting enables predictive rather than reactive scheduling (Maji et al., 2023). Once complete, training carbon can be amortized across all downstream applications.

Inference carbon accumulates continuously with each query served. Recent studies demonstrate that inference energy follows different scaling relationships than training (Desislavov et al., 2023; Samsi et al., 2023; Fernandez et al., 2025). Energy per token decreases with batch size but increases with sequence length. A large language model can consume several kWh per 1000 queries depending on model size, hardware, and configuration (Luccioni et al., 2024). For frontier models like GPT-3 (Brown et al., 2020) or Llama (Touvron et al., 2023a;b) deployed at massive scale, inference emissions can exceed training emissions within weeks or months (Chien et al., 2024).

This distinction has critical implications. Research papers typically report only training costs, but commercially deployed models incur ongoing inference costs that dwarf training investments. Papers should report training carbon with methodology details, and model releases should include standardized inference energy benchmarks to enable lifecycle carbon assessment.

Inference benchmarks must be domain-specific. For example, language models might report energy per 1000 tokens at specified batch sizes, while vision models might use energy per 1000 images. Classical ML, recommendation systems (Wegmeth et al., 2025), and AutoML have different profiles where per-sample metrics may be more appropriate. The key principle is standardization within each domain to enable fair comparison, while acknowledging that production deployments involve additional complexity that simplified benchmarks do not fully capture.

4.5. Measurement Protocol

To enable meaningful comparisons across papers, we propose a minimal viable protocol that specifies not just which metrics to report, but how to measure them. This protocol prioritizes comparability and transparency while acknowledging practical constraints.

How to measure. Report average power draw during the measurement period, not peak power or TDP, as nameplate ratings can overestimate by 20–40%. When using software-based tools, specify the sampling interval (e.g., 1-second intervals for CodeCarbon); for hardware power meters, specify the measurement point (GPU only, server, or rack-level). Software-based tools have known limitations, so report confidence intervals rather than false precision (e.g., "approximately 150 kWh $\pm$20%" rather than "153.7 kWh"). When possible, cross-validate estimates using multiple tools.

What to include. Clearly define what computation is included. The "final reported experiments" boundary should include all training runs whose results appear in the paper, final evaluation runs, and any experiment-specific preprocessing. Pre-trained weights need not be counted, but authors should cite reported training emissions of base models when available. When results are averaged across n runs, report total energy ($n\times$ single-run), not per-run averages. Report hyperparameter search separately from final training: "Final training: X kWh; Hyperparameter search: Y kWh (Z configurations evaluated)."

How to report. For multi-location compute, report contributions separately with location-specific carbon intensities, then provide a total (e.g., "Location A (Quebec): 500 kWh, 10 $kgCO_2eq$; Location B (Germany): 300 kWh, 105 $kgCO_2eq$; Total: 800 kWh, 115 $kgCO_2eq$"). Include normalized metrics (carbon per accuracy point, per training sample, or per parameter) alongside totals to enable fair comparison across different scales.

5. Call to Action

We call on the ML community to take concrete steps toward normalizing carbon footprint reporting.

5.1. Venues and Reviewers

Conference organizers hold the key to widespread adoption. We urge major venues such as NeurIPS, ICML, and ICLR to add optional carbon reporting fields to their submission forms within the next submission cycle, asking for hardware, location, energy consumption, and carbon emissions. Venues should commission best-practice guides that help authors understand what to report and how. Over subsequent stages, venues should transition from optional to encouraged reporting, giving the community time to adapt while signaling clear expectations. Aggregate statistics on community-wide carbon footprint should be published annually, enabling the field to track its environmental trajectory.

Reviewers must understand that carbon reporting serves transparency, not gatekeeping. Papers should never be penalized for reporting high carbon costs; doing so would create perverse incentives to underreport or omit data entirely. Scrutiny should focus on two cases: (1) efficiency claims without energy data, and (2) massive compute yielding marginal gains over simpler baselines. High carbon is acceptable when justified; the goal is informed evaluation, not carbon-based rejection.

5.2. Researchers and Institutions

Authors need not wait for venue requirements. Voluntary adoption builds community norms. When reporting, authors should document measurement methodology clearly, including tool versions, hardware specifications, and data center locations. Research teams can begin by integrating measurement tools into their training pipelines and establishing lab-level reporting practices that normalize carbon awareness within their groups.

Universities, national labs, and cloud providers can facilitate reporting by providing energy monitoring infrastructure and publishing PUE ratios and energy sources. Some institutions have begun offering carbon-aware scheduling, routing jobs to times when grid carbon intensity is lower. Institutional support is particularly important for researchers who lack direct access to power monitoring hardware; centralized infrastructure can provide the data that individual researchers cannot easily obtain.

5.3. Preventing Gaming and Strategic Underreporting

Any reporting requirement creates incentives for strategic behavior. We identify two categories of potential gaming and propose mitigations.

Selective reporting scope. Authors might cherry-pick which runs to include, reporting energy only for the best-performing run while averaging accuracy across multiple runs. Similarly, authors might define experiment boundaries to exclude expensive hyperparameter sweeps or omit preprocessing, data loading, and checkpointing from measurements. The measurement protocol addresses these concerns by requiring total energy across all averaged runs, separate reporting of search and training compute, and explicit documentation of what is excluded and why.

Carbon accounting manipulation. Authors using renewable energy credits (RECs)[8] or carbon offsets might report only net emissions, obscuring gross energy consumption. Following GHG Protocol guidance, we require reporting of both gross and net emissions, with energy consumption in kWh reported independently. A related concern is location arbitrage, where authors route computation through low-carbon regions primarily for reporting optics. However, this is actually a positive outcome: if reporting incentivizes using cleaner grids, that represents genuine emissions reduction.

We acknowledge that no reporting system is gaming-proof, but establishing clear norms makes egregious underreporting socially costly and statistically detectable.

5.4. Timeline for Adoption

We propose a phased transition in three stages, modeled on successful precedents: the shift from optional to expected code release, and the rapid adoption of NeurIPS

[8]Tradable certificates representing proof that electricity was generated from renewable sources.

reproducibility checklists (Pineau et al., 2021). The timeline is deliberately gradual: community norms shift slowly, and rushing requirements risks backlash that could set back adoption.

Stage 1: Foundation. Major venues add optional carbon reporting fields to submission forms, asking for GPU-hours, hardware specifications (GPU model, count), and data center location if known. Reviewers check only that reported data is present and internally consistent—for example, 8 A100 GPUs for 24 hours at typical utilization should yield roughly 50–150 kWh depending on hardware variant and workload, and gross inconsistencies (e.g., order-of-magnitude errors) warrant clarification but not rejection. Venues host workshops and tutorials on measurement tools, and early adopters share case studies documenting their reporting experience.

Stage 2: Encouragement. Venues publish aggregate statistics from submitted papers, establishing empirical ranges for different experiment types (e.g., fine-tuning a 7B model typically consumes 50–200 kWh; training a vision transformer from scratch typically consumes 500–2000 kWh). The expected requirement expands to measured energy (kWh) with acknowledged uncertainty bounds. Community benchmarks such as MLPerf incorporate efficiency metrics alongside accuracy. Third-party organizations including the Green Software Foundation offer voluntary verification services. Cross-venue data sharing agreements enable longitudinal analysis, and institutions begin offering carbon-aware job scheduling.

Stage 3: Normalization. Reporting becomes a community expectation rather than an exception. For large-scale experiments (exceeding thresholds such as 1000 GPU-hours or 10,000 kWh), full reporting of all five metrics becomes standard: kWh, $kgCO_2eq$, grid intensity, region, and PUE. Independent auditing frameworks emerge for frontier model training, potentially aligned with ISO 14064 certification standards. Venues recognize verified reports with badges similar to artifact evaluation, and carbon-aware practices integrate into graduate curricula.

Success at each stage is measured by concrete metrics: (1) *Adoption rate*—targeting 30% of papers including basic data (Stage 1), 50% including measured energy (Stage 2), and 70% meeting full requirements for applicable papers (Stage 3); (2) *Data quality*—fraction of reports that are internally consistent and include uncertainty estimates; (3) *Trend visibility*—ability to answer questions like "Is NLP research becoming more carbon-efficient?" with statistical confidence.

6. Alternative Views

A proposal of this scope raises legitimate concerns. Engaging with these objections strengthens our argument by forcing us to refine implementation details and acknowledge real trade-offs. We address four common objections to standardized carbon reporting.

6.1. Measurement Is Too Difficult

Critics argue that accurate energy measurement requires specialized hardware or software configurations that many researchers lack. Different tools produce different estimates, and the resulting numbers may be unreliable.

We acknowledge this concern. Estimates from software tools can vary by 20–30% from ground truth (Jay et al., 2023), and different tools exhibit systematic biases (e.g., CodeCarbon tends to underestimate compared to hardware meters). We therefore recommend reporting the measurement tool and version used, enabling future cross-calibration studies. However, the barrier to entry is lower than commonly perceived: tools like CodeCarbon require only a pip install, major cloud providers now offer built-in carbon dashboards, and many HPC clusters already log energy consumption at the job level through SLURM or similar schedulers. Community resources including tutorials, documentation, and worked examples continue to grow. We also advocate for tiered requirements: at minimum GPU-hours and hardware specifications, ideally measured kWh, with full carbon accounting ($kgCO_2eq$) where feasible.

Imperfect measurement is better than no measurement. Astronomy and climate science routinely report observations with uncertainty ranges. Even with 20–30% error margins, energy data reveals order-of-magnitude differences and enables trend analysis across the field.

6.2. Reporting Creates Barriers for Under-Resourced Groups

A more serious concern is equity. Researchers at well-funded institutions with efficient data centers will report lower carbon footprints than those at smaller institutions using older hardware or dirtier grids. Reporting expectations could disadvantage already-marginalized researchers.

We take this concern seriously and acknowledge a tension in our proposal: while we recommend against using carbon as a review criterion (Section 5.1), we also suggest that benchmarks incorporate carbon metrics (Section 4.2). This creates a foreseeable "carbon leaderboard" effect where researchers in high-carbon regions face reputational disadvantage even without formal acceptance penalties.

We propose three mechanisms to address this tension:

Dual reporting of energy and carbon. Leaderboards and benchmarks should display both energy consumption (kWh) and carbon emissions ($kgCO_2eq$) separately. Energy reflects algorithmic and hardware efficiency independent of

location; carbon reflects the full environmental impact. Researchers in high-carbon regions can demonstrate strong energy efficiency even if their carbon numbers are higher, and the community can recognize that geographic constraints differ from engineering choices.

Reference-grid normalization. For comparative rankings, venues could report carbon emissions normalized to a standard reference grid intensity (e.g., the global average of approximately 475 gCO_2/kWh). This "what-if" metric answers: "What would this experiment emit on an average grid?" Actual emissions should still be reported for transparency, but normalized values enable fairer cross-institution comparison of algorithmic efficiency.

Efficiency ratios over absolute values. Metrics like kWh per accuracy point or gCO_2eq per billion parameters reward efficiency regardless of scale. A researcher with limited compute who achieves high accuracy-per-joule demonstrates scientific craftsmanship that absolute carbon totals would obscure.

For researchers in the Global South or at institutions with minimal infrastructure support, we recommend practical accommodations: software-based tools require only pip install; approximate reporting (e.g., GPU-hours with estimated regional intensity) should be accepted when precise measurement is infeasible. The goal is inclusion, not exclusion.

Reporting should never be used as a gatekeeping criterion for acceptance. A breakthrough result remains valuable regardless of its carbon cost; what matters is that the cost is known. Our goal is informed decision-making, not restricting ambitious research—transparency about costs enables the community to weigh trade-offs explicitly rather than ignoring them.

6.3. Standards Are Premature

Some argue the field lacks consensus on measurement methods, and standardizing now will lock in imperfect approaches.

We disagree. Perfect standards are not required for useful reporting. The current state, where most papers report nothing, prevents progress on understanding the field's environmental trajectory. Even imperfect data enables trend analysis. Moreover, standardization drives improvement: once venues encourage reporting, tool developers will have stronger incentives to improve accuracy, and the community will converge on best practices through experience.

6.4. Focus Should Be on Industry, Not Academia

A fourth objection holds that academic research contributes a small fraction of total ML compute. The real environmental impact comes from industry training runs and deployed inference at scale. Requiring academics to report while industry operates opaquely is asymmetric and ineffective.

We agree that industry must be part of the solution. However, academic norms influence industry practice: researchers who learn to report environmental impact in graduate school carry these practices into industry roles. Academic venues can require that industry-sponsored publications meet the same reporting standards as academic submissions, creating pressure for organizational transparency. Moreover, as venues publish aggregate statistics, industry papers that report nothing will become conspicuous outliers.

The objection also underestimates academic impact. While individual academic experiments may be small, the aggregate compute across thousands of papers is substantial. More importantly, academic research shapes what problems the field considers important. If efficiency and environmental impact become standard metrics in academic publications, industry researchers will face pressure to adopt similar standards.

7. Conclusion

The ML community has developed rigorous standards for reporting methodology, statistics, and compute requirements. Carbon footprint reporting is the logical next step. Our argument is scientific: efficiency claims without energy data are incomplete, analogous to reporting accuracy without specifying the test set. Systematic reporting would let us answer questions we currently cannot: Is ML research becoming more carbon-efficient over time? Do algorithmic gains translate to real energy savings? With consistent reporting, we could establish baselines, track progress, and identify where intervention is most needed.

We acknowledge challenges, but the field has adopted comparably difficult norms before: code release evolved from rare to expected within a decade; NeurIPS introduced reproducibility checklists in 2019 and achieved widespread adoption within two years (Pineau et al., 2021); Datasheets for Datasets (Gebru et al., 2021) moved from proposal to common practice. In each case, initial resistance gave way to recognition that transparency benefits outweigh costs. Reporting must be implemented with equity in mind: dual energy-carbon reporting, reference-grid normalization, and acceptance of approximate estimates aim to make reporting inclusive rather than exclusionary.

The path forward requires coordination: venues must signal that reporting is valued, authors must adopt measurement tools, and institutions must provide infrastructure support. The tools exist. The precedents exist. This norm is overdue.

References

Anthony, L. F. W., Kanding, B., and Selvan, R. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. In *ICML Workshop on Challenges in Deploying and Monitoring Machine Learning Systems*, 2020.

Bian, Z., Wang, L., and Ren, S. CAFE: Carbon-aware federated learning in geographically distributed data centers. In *IEEE International Conference on Computer Communications (INFOCOM)*, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Chien, A. A., Lin, L., and Nguyen, H. Reducing the carbon impact of generative AI inference (today and in 2035). *arXiv preprint arXiv:2409.02839*, 2024.

CodeCarbon Contributors. CodeCarbon: Track and reduce CO2 emissions from your computing. https://codecarbon.io/, 2020.

Desislavov, R., Martinez, F., Sherlock, M., and Sherlock, M. Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems*, 38:100857, 2023.

Dodge, J., Gururangan, S., Card, D., Schwartz, R., and Smith, N. A. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194, 2019.

Dodge, J., Prewitt, T., Des Combes, R. T., Buchanan, E., Duber, T., Ganguli, D., et al. Measuring the carbon intensity of AI in cloud instances. In *ACM Conference on Fairness, Accountability, and Transparency*, pp. 1877–1894, 2022.

Electricity Maps. Real-time carbon intensity data for global power grids. https://www.electricitymaps.com/, 2023.

Fernandez, J., Na, C., Tiwari, V., Bisk, Y., Luccioni, S., and Strubell, E. Energy considerations of large language model inference and efficiency optimizations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., et al. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Green Software Foundation. Software carbon intensity (SCI) specification. *Green Software Foundation Standard*, 2022.

Gupta, U., Kim, Y. G., Lee, S., Tse, J., Lee, H.-H. S., Wei, G.-Y., et al. Chasing carbon: The elusive environmental footprint of computing. *IEEE Micro*, 42(4):37–47, 2022.

Hasan, S. M., Shahid, A. R., and Imteaj, A. Towards sustainable SecureML: Quantifying carbon footprint of adversarial machine learning. *arXiv preprint arXiv:2401.08577*, 2024.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Jay, M., Ostapenco, V., Lefèvre, L., Trystram, D., Orgerie, A.-C., and Fichel, B. An experimental comparison of software-based power meters: Focus on CPU and GPU. In *IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pp. 106–118, 2023.

Jegham, N., Abdelatti, M., Koh, C. Y., Elmoubarki, L., and Hendawi, A. How hungry is AI? benchmarking energy, water, and carbon footprint of LLM inference. *arXiv preprint arXiv:2505.09598*, 2025.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

K.C., A. et al. Towards eco-friendly cybersecurity: Machine learning based anomaly detection with carbon and energy metrics. *arXiv preprint arXiv:2601.00893*, 2025.

Knödlseder, J., Brau-Nogué, S., Coriat, M., Garnier, P., Hughes, A., Martin, P., and Tibaldo, L. Estimate of the carbon footprint of astronomical research infrastructures. *Nature Astronomy*, 6:503–513, 2022.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. Quantifying the carbon emissions of machine learning. In *NeurIPS Workshop on Tackling Climate Change with Machine Learning*, 2019.

Li, B., Samsi, S., Gadepally, V., and Tiwari, D. Clover: Toward sustainable AI with carbon-aware machine learning inference service. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2023.

Ligozat, A.-L., Lefèvre, J., Bugeau, A., and Combaz, J. Unraveling the hidden environmental impacts of AI solutions for environment life cycle assessment of AI solutions. *Sustainability*, 14(9):5172, 2022.

Lottick, K., Susai, S., and Friedler, S. A. Energy usage reports: Environmental awareness as part of algorithmic accountability. In *Workshop on Tackling Climate Change with Machine Learning at NeurIPS*, 2019.

Luccioni, A. S., Viguier, S., and Ligozat, A.-L. Estimating the carbon footprint of BLOOM, a 176b parameter language model. *Journal of Machine Learning Research*, 24 (253):1–15, 2023.

Luccioni, A. S., Jernite, Y., and Strubell, E. Power hungry processing: Watts driving the cost of AI deployment? *arXiv preprint arXiv:2311.16863*, 2024.

Maji, D., Shenoy, P., and Sitaraman, R. K. Multi-day forecasting of electric grid carbon intensity using machine learning. In *Proceedings of the 14th ACM International Conference on Future Energy Systems (e-Energy)*, pp. 314–325, 2023.

Masanet, E., Shehabi, A., Lei, N., Smith, S., and Koomey, J. Recalibrating global data center energy-use estimates. *Science*, 367(6481):984–986, 2020.

Meta AI. Llama 3.1 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md, 2024.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., et al. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., et al. Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, 22(164):1–20, 2021.

Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., et al. From words to watts: Benchmarking the energy costs of large language model inference. *arXiv preprint arXiv:2310.03003*, 2023.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. Green AI. *Communications of the ACM*, 63(12):54–63, 2020.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. Compute trends across three eras of machine learning. *arXiv preprint arXiv:2202.05924*, 2022.

Strubell, E., Ganesh, A., and McCallum, A. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650, 2019.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Tschand, A., Rajan, A., Kuppannagari, S., Kanter, D., and Reddi, V. J. MLPerf power: Benchmarking the energy efficiency of machine learning systems from microwatts to megawatts for sustainable AI. In *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2025.

Verdecchia, R., Lago, P., and de Vries, J. Green AI techniques for reducing energy consumption in AI systems. *Journal of Systems and Software*, 2025.

Wegmeth, L., Vente, T., and Said, A. Green recommender systems: Understanding and minimizing the carbon footprint of AI-powered personalization. *arXiv preprint arXiv:2501.03988*, 2025.

World Resources Institute and World Business Council for Sustainable Development. The greenhouse gas protocol: A corporate accounting and reporting standard. https://ghgprotocol.org/corporate-standard, 2004.

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Arber, N., Cho, K., et al. Sustainable AI: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813, 2022.

Xiao, T., Fuso Nerini, F., Matthews, H. D., Tavoni, M., and You, F. Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the USA. *Nature Sustainability*, 8:1528–1540, 2025.

Yu, Z., Wu, Y., Deng, Z., Yan, Y., Jia, Y., Zheng, B., and Ren, Q. OpenCarbonEval: A unified carbon emission estimation framework in large-scale AI models. *arXiv preprint arXiv:2405.12049*, 2024.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A. Glossary of Key Terms

Table 1. Glossary of carbon footprint reporting terminology.

Term	Definition
Carbon Footprint	The total greenhouse gas emissions caused directly and indirectly by an activity, expressed as carbon dioxide equivalent (CO_2eq).
Carbon Intensity	The amount of CO_2 emitted per unit of electricity generated, typically measured in gCO_2/kWh. Varies by region and time based on energy mix.
CO_2eq	Carbon dioxide equivalent; a standard unit for measuring carbon footprints that converts all greenhouse gases to the equivalent amount of CO_2 based on global warming potential.
GPU-hours	A measure of computational work equal to one GPU operating for one hour. Does not account for GPU utilization or power draw variations.
PUE	Power Usage Effectiveness; the ratio of total facility energy to IT equipment energy. A PUE of 1.0 means all power goes to computing; typical data centers range from 1.1 to 2.0.
Scope 1 Emissions	Direct emissions from owned or controlled sources (e.g., on-site generators).
Scope 2 Emissions	Indirect emissions from purchased electricity, steam, heating, and cooling. Most ML carbon footprints fall here.
Scope 3 Emissions	All other indirect emissions in the value chain, including hardware manufacturing and end-of-life disposal.
TDP	Thermal Design Power; the maximum amount of heat a component is designed to dissipate, often used as a proxy for power consumption.
FLOPs	Floating-point operations; a measure of computational work independent of hardware or energy consumption.
Embodied Carbon	The carbon emissions associated with manufacturing, transporting, and disposing of hardware, distinct from operational emissions.

B. Sample Carbon Reporting Templates

Table 2. Minimal reporting (Stage 1): GPU-hours and hardware only.

Metric	Value
Hardware	2× NVIDIA RTX 3090
Total GPU-hours	48
Compute region	University cluster (unknown grid)
Energy/carbon not measured; estimated ∼15 kWh based on TDP.	

C. Reference Data

Formulas: Total energy $= E_{\text{compute}} \times$ PUE (compute energy $\times$ data center overhead). Carbon footprint $= E_{\text{total}} \times I$ (total energy $\times$ grid intensity). Energy estimate $=$ GPU-hours $\times P_{\text{GPU}} \times U +$ overhead (GPU power at 70–80% TDP, plus 10–20% for CPU/memory).

Table 3. Standard reporting (Stage 2): includes measured energy. *Illustrative example; values are hypothetical.*

Metric	**Value**
Hardware	8× NVIDIA A100 (40GB)
Total GPU-hours	2,400
Compute region	US-West (California)
Grid carbon intensity	210 gCO_2/kWh
Data center PUE	1.1
Total energy (estimated)	1,000 kWh
Total carbon (estimated)	210 $kgCO_2eq$
Measurement tool	CodeCarbon v2.1

Table 4. Comprehensive reporting (Stage 3): full carbon accounting for large-scale experiments. *Illustrative example; values are hypothetical.*

Category	**Details**
Hardware Configuration	
GPU Type	NVIDIA H100 (80GB)
Number of GPUs	256
CPU	AMD EPYC 7763 (per node)
Interconnect	InfiniBand NDR 400Gb/s
Compute Summary	
Total GPU-hours	86,016 (256 GPUs × 14 days)
Average GPU Utilization	78%
Peak Power Draw (measured)	148 kW
Location & Infrastructure	
Data Center Location	Iowa, USA
Grid Carbon Intensity	380 gCO_2/kWh (annual average)
Renewable Energy Credits	50% offset claimed
PUE	1.08
Energy & Emissions	
Total Energy (measured)	62,000 kWh
Scope 2 Emissions (gross)	23,560 $kgCO_2eq$
Scope 2 Emissions (net, after RECs)	11,780 $kgCO_2eq$
Contextual Comparisons	
Equivalent car miles	59,000 miles
Equivalent transatlantic flights	13 round trips
Measurement Method	Direct power metering + CodeCarbon validation

Table 5. Grid carbon intensities by region (recent averages).[a]

Region	gCO_2/kWh	Cloud Zone
Quebec	20	GCP: na-northeast1
Norway	29	Azure: Norway East
France	60	GCP: europe-west9
UK	125	Azure: UK South
California	210	GCP: us-west1
Massachusetts	280	AWS: us-east-1
Germany	350	AWS: eu-central-1
US (average)	384	—
Global (average)	473	—
Poland	500	—
China	560	Alibaba: cn-hangzhou
Australia	650	GCP: au-southeast1
India	700	AWS: ap-south-1

[a]Sources: Electricity Maps (https://app.electricitymaps.com), IEA Emissions Factors 2025 (https://www.iea.org/data-and-statistics/data-product/emissions-factors-2025), Ember Global Electricity Review 2025 (https://ember-energy.org/latest-insights/global-electricity-review-2025), Carbon Brief (https://www.carbonbrief.org).

Table 6. Typical GPU power consumption during training.[b]

GPU	TDP	Typical Training
A10	150W	120W
V100 (32GB)	300W	250W
L40	300W	250W
RTX 3090	350W	290W
A100 (40GB)	400W	330W
A100 (80GB)	400W	350W
RTX 4090	450W	380W
MI250X	560W	470W
H100 (80GB)	700W	580W

[b]TDP from official NVIDIA/AMD specifications. Typical training power at 80–85% TDP. Sources: NVIDIA Data Center (https://www.nvidia.com/en-us/data-center), NVIDIA GeForce (https://www.nvidia.com/en-us/geforce).

D. Existing Tools and Standards

Table 7. Overview of carbon footprint measurement tools for ML research.

Tool	Measurement Method	Features & Limitations
CodeCarbon	Software-based power estimation using Intel RAPL[a] and nvidia-smi[b]	Easy integration with Python; cross-platform; may underestimate by 10–20% compared to hardware meters
Carbontracker	Software-based with hardware abstraction	Epoch-level tracking; predictive estimates; limited to Linux
experiment-impact-tracker	Software-based with regional carbon data	Comprehensive logging; JSON output; academic-focused
ML CO2 Impact	Web calculator using GPU-hours	Quick estimates; no code integration; uses average power values
Cloud Carbon Footprint	Cloud provider APIs	Works with AWS, GCP, Azure; relies on provider-reported data
Hardware-based Methods		
Power meters (e.g., Watts Up Pro)	Direct electrical measurement	Most accurate; requires physical access; measures at outlet level
IPMI/BMC sensors[c]	Server-level power reporting	Built into enterprise servers; 5–10% accuracy
PDU metering[d]	Rack-level power monitoring	Available in data centers; includes all equipment in rack

[a]Running Average Power Limit, Intel's interface for energy consumption monitoring. [b]NVIDIA System Management Interface, a command-line tool for querying GPU power draw. [c]Intelligent Platform Management Interface / Baseboard Management Controller. [d]Power Distribution Unit.

Table 8. Relevant standards and frameworks for carbon reporting.

Standard/Framework	Description & Relevance
GHG Protocol	The most widely used international standard for corporate carbon accounting. Defines Scope 1, 2, and 3 emissions categories.
ISO 14064	International standard for quantification and reporting of greenhouse gas emissions. Provides verification requirements.
ISO 14067	Standard specifically for carbon footprint of products, applicable to ML models as products.
Science Based Targets initiative (SBTi)	Framework for setting emission reduction targets aligned with climate science. Increasingly adopted by tech companies.
IEEE P2874	Proposed standard for carbon footprint metrics specifically for AI systems (under development).
Green Software Foundation	Industry consortium developing standards for sustainable software, including the Software Carbon Intensity (SCI) specification.

E. Worked Example

A research team fine-tunes a 7B model: 4× A100 GPUs, 48 hours, Massachusetts (280 gCO_2/kWh), PUE 1.4.

Calculation: GPU-hours = $4 \times 48 = 192$. GPU energy = 192×0.33 kW ≈ 63 kWh. With 15% overhead and PUE: $63 \times 1.15 \times 1.4 \approx 102$ kWh. Carbon $\approx 102 \times 0.28 \approx 29$ kgCO_2eq ($\pm$20%).

Report: $\approx$29 kgCO_2eq ($\pm$20%) $\approx$ 70 car miles. The same experiment in Quebec (20 gCO_2/kWh) would produce $\approx$2 kg; in Australia (650 gCO_2/kWh), $\approx$66 kg—a 33× variation.